EVALUACIÓN DE VISUALIZACIONES EFICIENTES EN CIENCIA DE DATOS

Mag. Raúl Oscar Klenzi, Mag. María Alejandra Malberti, Mag. Graciela Elida Beguerí

Instituto de Informática / Departamento de Informática /Facultad de Ciencias Exactas Físicas y Naturales / Universidad Nacional de San Juan

Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", Rivadavia, San Juan, Teléfonos: 4260353, 4260355 Fax 0264-4234980, Sitio Web: http://www.exactas.unsj.edu.ar
e-mail: {rauloscarklenzi,amalberti,grabeda}@gmail.com

RESUMEN

El presente proyecto plantea como objetivo proponer criterios para la evaluación de visualizaciones eficientes en Ciencia de Datos.
A tal fin se tiene pensado investigar sobre distintos aspectos que atañen a una visualización tales como escala, longitud, área, color, entre otros. También examinar herramientas libres con capacidades para visualización de datos e información y lenguajes de propósito general como Python y JavaScript.
De este modo se espera analizar la aptitud de diversas visualizaciones y sugerir métricas para su evaluación.

Palabras clave: Visualización, Ciencia de Datos, Lenguajes de código abierto, Software libre.

CONTEXTO

Los avances tecnológicos, tanto en potencialidades de cómputo, capacidades de almacenamiento, lenguajes de programación, herramientas de software, entre otros, permiten que más y más personas y empresas generen y guarden datos. Esta gran cantidad de datos almacenados requiere no solo de herramientas para procesarlos, sino también de "formas" de entenderlos y percibir rápidamente la información que contienen. Si bien la representación de la información tiene la capacidad de sintetizar y condensar datos, y constituye un enfoque eficiente para el análisis; no resuelve todos los problemas. (Erraissi y otros, 2019)

Si bien la visualización de la información no resuelve todos los problemas, es una parte importante en varias de las etapas de la Ciencia de datos.

La propuesta cuenta con antecedentes logrados en el tema conforme a los sucesivos proyectos aprobados y subsidiados por CICITCA - UNSJ en los que el grupo ha trabajado, siendo estos:

- "Búsqueda inteligente de información no relacionada en grandes bases de datos" 21/E508, período 2003-2005
- "Descubrimiento de conocimiento a través de Data Warehousing y Data Mining, en los datos de la Biblioteca de la Facultad de Ciencias Exactas, Físicas y Naturales" 21/E639, período 2005-2007
- "Búsqueda estratégica de conocimiento en los datos de biblioteca y de alumnos de la FCEFN" 21/E824, período 2008-2010
- "Minería de datos en la determinación de patrones de uso y perfiles de usuarios" 21/E889, período 2011-2013
- "Extracción de Conocimiento en Datos Masivos" 21/E-951, período 2014-2015
- "La Ciencia de Datos en grandes colecciones de datos" 21/E1014, período 2016-2017

- "Visualización y Deep Learning en Ciencia de Datos" 21/E1071, período 2018-2019

1. INTRODUCCIÓN

La Ciencia de Datos es una rama del paradigma de big data, un enfoque en el que se capturan, procesan a velocidades enormes, variedad y volúmenes de datos estructurados, no estructurados y semiestructurados, utilizando un conjunto de técnicas y tecnologías completamente novedosas en comparación con las que se utilizaron en décadas anteriores. Es útil para derivar conocimiento a partir de datos en bruto, esencial para cualquier sistema integral de apoyo a la toma de decisiones y extremadamente importante a la hora de formular estrategias sólidas para la gestión empresarial.
A la Ciencia de Datos se la puede pensar como el dominio científico dedicado al descubrimiento de conocimiento mediante análisis de datos. Es decir, se refiere al sector de la industria o dominio temático que los métodos de Ciencia de Datos utilizan para explorar. Los científicos de datos aplican técnicas matemáticas y enfoques algorítmicos para derivar soluciones a problemas complejos empresariales y científicos. Tanto en los negocios como en la ciencia, estos métodos pueden proporcionar capacidades de toma de decisiones más robustas.

"Visualización es aquella tecnología plural (esto es, disciplina) que consiste en transformar datos en información semántica — o en crear las herramientas para que cualquier persona complete por sí sola dicho proceso — por medio de una sintaxis de fronteras imprecisas y en constante evolución basada en la conjunción de signos de naturaleza icónica (figurativos) con otros de naturaleza arbitraria y abstracta (no figurativos: textos, estadísticas, etc.)". (Cairo, 2011)
En Barcellos y otros (2017) se argumenta que el uso de algunos tipos específicos de visualizaciones, asociados con el análisis de datos y las técnicas de aprendizaje automático, pueden ayudar en la interpretación de aspectos univariados, bivariados y multivariados de los datos.
Al diseñar visualizaciones cuanto más se conozca a los destinatarios, mejor posicionado se estará para comprender cómo resonar en ellos y así formar una comunicación que satisfaga sus necesidades. Comprender el contexto, elegir una pantalla visual adecuada, eliminar el desorden, decidir dónde dirigir la atención, deben ser tenidos en cuenta al diseñar y crear visualizaciones para impartir información. (Knaflic, 2015)

Desde hace varios años KNIME Analytics (herramienta de software libre) permite incorporar, en forma modular, todos los pasos involucrados en el proceso de Ciencia de Datos. Esta herramienta proporciona un repositorio de módulos fáciles de usar. Adicionalmente a las técnicas estándares de Minería de Datos, añade los algoritmos más actuales de análisis tales como los de deep learning. Posee integraciones con Python, JavaScript, R y con otros grandes proyectos de código abierto, dando libertad de mezclar y combinar las herramientas que se deseen, dentro de un entorno uniforme. (Klenzi, R y otros 2019)

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Entre las líneas más importantes de investigación y desarrollo se mencionan:

- Ciencia de Datos, principalmente Visualización de Información
- Deep Learning
- Herramientas de software libre para arquitecturas secuenciales, paralelas y distribuidas
- Lenguajes de programación de código abierto tales como Python y JavaScript.

3. RESULTADOS OBTENIDOS/ESPERADOS

Se tiene como objetivo proponer criterios para la evaluación de visualizaciones eficientes en Ciencia de Datos.

También se pretende:

- Analizar distintos aspectos que atañen a una representación visual
- Realizar búsqueda de datos abiertos y recopilar otras fuentes de datos provenientes de actividades de cooperación
- Evaluar herramientas de software libre en cuanto a sus potencialidades de visualización y eficiencia computacional
- Realizar análisis comparativo de los lenguajes Python y JavaScript como soporte de visualizaciones
- Seleccionar visualizaciones y clasificarlas de acuerdo a tipos de datos que admiten, usuarios a las que estén dirigidas, así como lenguajes y herramientas que las soportan.

4. FORMACIÓN DE RECURSOS HUMANOS

El equipo de docentes investigadores del proyecto trabaja desde hace varios años en diversas áreas relacionadas con la Ciencia de Datos.
Es de destacar también, que la propuesta incorpora ayudantes y alumnos que cursan diferentes carreras de FCEFN-UNSJ. Es convencimiento de los integrantes del proyecto que la formación de estos jóvenes en el área de la investigación y áreas temáticas inherentes a la Ciencia de Datos enriquecerá las diferentes actividades y son de suma importancia para su futuro profesional y laboral.
Actualmente forman parte del equipo los coordinadores del “Laboratorio de Sistemas Inteligentes para la Búsqueda de Datos” del Instituto de Informática-FCEFN y los docentes titulares de las asignaturas Inteligencia Artificial, Probabilidad y Estadística, Sistemas de Datos, Estructuras de Datos y Algoritmos, Programación Procedural y Programación Orientada a Objetos, de las carreras Licenciatura en Sistemas de Información y Licenciatura en Ciencias de la Computación pertenecientes al Departamento de Informática de la FCEFN-UNSJ.
Se habrán de proponer en el bienio de duración del proyecto, trabajos finales de grado, posgrado y becas de investigación.
En particular, con docentes investigadores del INAUT-FI y personal del INTA-San Juan, se llevará adelante la tesis de maestría de un integrante del equipo de trabajo "Análisis de Fenómenos en Estaciones Agrometeorológicas Mediante Ciencia de Datos".

Equipo de investigación:

Director: Mag. Raúl Oscar Klenzi
Co-Director: Mag. María Alejandra Malberti Riveros

Integrantes docentes-investigadores:

- Mag. Graciela Elida Beguerí
- Lic. Laura Gutiérrez
- Lic. María Isabel Masanet Yáñez
- Lic. Manuel Ortega
- Prog. Luis Olguín
- Lic. Fabrizio Amaya

Integrantes alumnos:

- Joaquín Cortez
- Carolina Olivares
- Malena Páez
- Noelia Pérez

5. BIBLIOGRAFÍA

- Barcellos, R., Viterbo, J., Miranda, L., Bernardini, F., Maciel, C., & Trevisan, D. (2017, June). Transparency in practice: using visualization to

enhance the interpretability of open data. In Proceedings of the 18th Annual International Conference on Digital Government Research (pp. 139-148). ACM.

- Barcellos, R., Viterbo, J., Bernardini, F., & Trevisan, D. (2018, July). An Instrument for Evaluating the Quality of Data Visualizations. In 2018 22nd International Conference Information Visualisation (IV) (pp. 169-174). IEEE.
- Benoit, G. (2019). Introduction to Information Visualization: Transforming Data Into Meaningful Information. Rowman & Littlefield.
- Cady, F. (2017). The data science handbook. John Wiley & Sons.
- Cairo, A. (2011). El arte funcional: infografía y visualización de información. Alamut.
- Castro Almudena M. e Úcar Iñaki (2019). Gráficas para la ciencia y ciencia para las gráficas. Cuaderno de Cultura Científica. https://culturacientifica.com/2019/04/25/graficas-para-la-ciencia-y-ciencia-para-las-graficas/
- Chen, M., Feixas, M., Viola, I., Bardera, A., Shen, H. W., & Sbert, M. (2017). Information theory tools for visualization. AK Peters/CRC Press
- Cielen, D., Meysman, A., & Ali, M. (2016). Introducing data science: big data, machine learning, and more, using Python tools. Manning Publications Co..
- Erraissi, A., Mouad, B., & Belangour, A. (2019, April). A Big Data visualization layer meta-model proposition. In 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO) (pp. 1-5). IEEE.
- Grus, J. (2019). Data science from scratch: first principles with python. O'Reilly Media.
- How to choose a visualization (2019). https://www.kdnuggets.com/2019/06/how-choose-visualization.html
- Irisarri Patricio (2017, Junio). Breve historia de la visualización de datos. http://periodismodigitaliset18.blogspot.com/2017/06/breve-historia-de-la-visualizacion-de.html
- Klenzi, R. O., Malberti, A., & Beguerí, G. (2019, June). Propuesta didáctica inherente al área de ciencia de datos. In *XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan).*
- Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons.
- Laura Igual Muñoz, & Santi Seguí Mesquida. (2017). Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications. Springer.
- Lo, L. Y. H., Ming, Y., & Qu, H. (2019). Learning Vis Tools: Teaching Data Visualization Tutorials. arXiv preprint arXiv:1907.08796.
- Murray, S. (2017). Interactive data visualization for the web: an introduction to designing with D3" O'Reilly Media, Inc.".
- Schutt, R., & O'Neil, C. (2013). Doing data science: Straight talk from the frontline. O'Reilly Media, Inc.
- Schwabish Jonathan A. (2014, February). A Visualization Mapping: Form and Function

https://policyviz.com/2014/02/05/a-visualization-mapping-form-and-function/

- Sosulski, K. (2018). Data Visualization Made Simple: Insights Into Becoming Visual. Routledge.
- Tufte, E. R. (2001). The visual display of quantitative information (Vol. 2). Cheshire, CT: Graphics press.
- VanderPlas, J. (2016). Python data science handbook: essential tools for working with data. " O'Reilly Media, Inc.".
- Wang, J., Hazarika, S., Li, C., & Shen, H. W. (2018). Visualization and visual analysis of ensemble data: A survey. IEEE transactions on visualization and computer graphics.
- Wang, C., & Shen, H. (2011). Information Theory in Scientific Visualization. Entropy, 13, 254-273.
- Ware, C. (2012). Information visualization: perception for design. Elsevier
- Wenqiang Cui (2019) Visual Analytics: A Comprehensive Overview. EEE Xplore Digital Library. Volume 7: 2019 https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8740868
- Why Data Visualization Is The Most Important Skill in a Data Analyst Arsenal (2019). https://www.kdnuggets.com/2019/08/simpliv-data-visualization-data-analyst.html
- Wilke, C. O. (2019). Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures. O'Reilly Media.
- Zuk, T., Schlesier, L., Neumann, P., Hancock, M. S., & Carpendale, S. (2006, May). Heuristics for information visualization evaluation. In Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization (pp. 1-6). ACM.